Introduction

This volume is about *computational phraseology*, a fairly recent notion (Colson, 2003; Granger & Meunier, 2008; Heid, 2008). While *computational linguistics* as a whole has become one of the main research fields of linguistics, it is also worthy of note that recent years have seen many *computational* subdisciplines gaining some ground. At the time of writing, *computational sociolinguistics* yields 3,400 hits[1] on Google, *computational psycholinguistics* 32,400, computational *discourse analysis* 1,660 and *computational construction grammar* 3,690. *Computational phraseology* is also progressing, with almost 1,000 hits. There must be some logic to this more frequent use of the adjective *computational* in various fields.

Our point of view is that part of the explanation may be related to the *people* who carry out computational research in the various subdisciplines of linguistics. Their background can be very different: linguists, but also engineers, mathematicians, statisticians, computer scientists, programmers, etc. Bringing very diverse people together for research activities is a daunting challenge. Engineers, for instance, rely a lot on statistics and may not be aware of the terminology and cognitive aspects of linguistics; linguists, on the other hand, do not always realise that improving automated linguistic techniques also requires some mastery of computer science or statistics. The success of the *computational* subdisciplines of linguistics may precisely be due to the fact that they foster collaboration between linguists on the one hand, and engineers / statisticians / computer scientists on the other. These approaches work *bottom-up*: the research questions and the terminology come from raw linguistic data, and are then investigated within a collaborative project.

This is also the case for phraseology, the study of all fixed multiword expressions, from collocations and formulas to proverbs. However, *computational phraseology* is not yet very common, as mentioned above. There may be historical reasons for this. Phraseology as a discipline is mainly represented by Europhras[2], the European association for phraseology, with a long research tradition coming from mainland Europe, and in particular from Russia and the German-speaking countries. Continental phraseology grew out of traditional linguistic approaches that focus on certain features of *phraseological units* or *phrasemes* (basically stability, idiomaticiy and gradability), and it has only recently started using automated techniques. On the other hand, the NLP and Computational Linguistics community have mainly focussed on the polylexical nature of these units, with a clear preference for terms like *multiword expression* (MWE), *multiword unit* or and *polylexical expression*. In this tradition, the focus has been on automatic identification, extraction and processing of MWE, with little or no reference to other linguistic features, apart from idiomaticity (Monti et al., 2018). Two worlds apart...

On the other hand, phraseology also has close historical links to corpus linguistics: Sinclair (1991) lays stress on the importance of the *idiom principle*, according to which roughly 50 percent of any text consists of phraseology in the broad sense. More recent studies carried out within the framework of corpus linguistics have explored other aspects of phraseology in various corpora of L1 or L2 users (Granger and Meunier, 2008). Our purpose here is not to expand on the subtle differences between computational linguistics and corpus linguistics, but one of the reasons why *computational phraseology* is relatively infrequent a notion may be due to the competition with alternative terms referring to corpora, such as *phraseology and corpora* or *corpus-based phraseology*.

---

[1] Google.com, last consulted on 23 May 2018. The search term used the quotation marks.
[2] www.europhras.org

As a matter of fact, many researchers from computational and corpus linguistics were actually dealing with phraseology without using this term. This certainly holds true for more than 50 years research on collocations and their automatic extraction from corpora (Gries, 2013). While *collocation* is sometimes used by computational linguists in the general sense of fixed expression, there is now a broad consensus as to the position of *collocations* at the left-hand side of a continuum ranging from weakly idiomatic expressions to idioms and proverbs. In other words, studying collocations is hardly possible without taking into account the whole spectrum of weakly idiomatic / fixed and idiomatic / very fixed and highly idiomatic expressions. Besides, a broad array of phraseological studies deal with the complex interweaving between idiomaticity, language and culture. To give just one example, the Chinese 4-character idiomatic expressions known as *chéngyǔ* (成语) only make sense with reference to Chinese culture. They indeed correspond to an older period of the language, are often linked to literature and are fixed in the linguistic competence of native speakers of Chinese as sequences consisting (in principle) of just 4 characters.

It doesn't take a specialist in phraseology to realise that most idiomatic expressions in any language are thus part of a complex network of cultural and linguistic elements. For English, the high proportion of phrases of maritime origin comes to mind, but more subtle links with culture or history can also be traced down. The popular phrase *be over the moon*, for instance, finds its origin in English nursery rhymes from the 18th century (Oxford Dictionary, online edition 2018), which underlines one particular literary genre and an aspect of British culture that had a major impact on English phraseology.

Against the background of history, society and culture, fixed associations in language can only be studied efficiently if a more general perspective is taken, which is precisely one of the goals of phraseology. Likewise, the notion of *computational phraseology* has the advantage of applying automated or corpus-based approaches to both linguistic and socio-cultural associations. This approach is also compatible with the recent developments of *construction grammar*, which sees language as a complex network of constructions, i.e. of pairings of form and meaning at different levels of abstraction and schematicity, and in relation with the culture of a specific language.

Let us take the example of the partly schematic *All-cleft* Construction, as in *All I had to do was find the correct answer*. This construction displays a complex network of schematic (i.e. interchangeable) and specific slots, and requires at least the pronoun *All* and two verbs as fixed slots: *All NP/PRO VP VP*. Corpus results show that a personal pronoun often follows the first slot (*All I had to go upon was…*), but other features display some regularity and might therefore be captured by statistical metrics: *All he achieved was, All he did the whole time was, All he had really expected was…*, etc. It is also clear that this construction can at any time head towards the phraseological end of the spectrum and yield a cliché or formula, as in *All you need is love*. The interplay with culture and society is clear in the last example, as well as the fuzzy border between partly fixed or partly schematic constructions and phraseology. Approaching such phenomena from the point of view of computational phraseology has the advantage of allowing for socio-cultural elements in the description, and of being compatible with a more general perspective on language.

In this book, the various aspects of the interplay between corpora, automated approaches and phraseology are illustrated.

In the first chapter, František Čermák discusses the results of a corpus-based, multilingual investigation into a special category of phrasemes that has received little attention, viz. *monocollocable words,* i.e. words that are so restricted in their combinations that they (almost)

only occur in a limited numbers of phrasemes, e.g. the word *ado* in *much ado about nothing*. Interestingly, such a phenomenon turns out to be present in many languages, but extracting relevant examples poses many problems to corpus-based or computational linguistics. In this contribution, a first selection of relevant cases by means of a statistical method had to be completed with manual annotation.

Another challenge for computational phraseology is the improvement of machine translation for phrasemes or *multiword expressions* (MWEs). In the next chapter, Johanna Monti, Mihael Arcan and Federico Sangati present an original way of tackling this thorny issue, by investigating the translation asymmetries of MWEs between English and Italian in corpora. For instance, an idiom such as *spill the beans* may be translated in another language by just one word, or vice versa. This has a major impact on the quality of statistical machine translation (SMT), as the latter largely relies on translation corpora. The originality of this contribution is to rely on a MWE-annotated bilingual corpus, and to compare it with the results obtained by machine translation for MWEs, which opens up many possibilities for further research.

Another fascinating and recently investigated area of research for corpus-based phraseology is the interaction with constructions, as defined by construction grammar. In his contribution to this volume, Dmitrij Dobrovol'skij describes the German constructional idioms [vor sich her + V] and [vor sich hin + V] based on corpus examples. Constructions of this type are not only problematic for translation into another language, but they are particularly difficult to describe in dictionaries, in spite of their great importance in the language. This study shows what is at stake in the analysis and description of such borderline cases between phraseology and construction grammar, for which large linguistic corpora are of the essence. The author also pleads for a fruitful collaboration between phraseology and construction grammar, in order to shed light on the broad array of partly compositional constructions such as those under investigation.

In chapter four, Jean-Pierre Colson argues that a corpus-based and computational approach may shed some fresh light on the intertwining of phraseology, culture and translation. For instance, in spite of the largely similar and very frequent words of which they are made, the communicative phrasemes *That's life* and *This is the life* have a totally different meaning, which is to be situated against the backdrop of cultural elements, idiomaticity and recurrent patterns, and may create translation problems. Phraseology is a daunting challenge for human translators, as they have to decode very accurately all idiomatic meanings in the source text, and look for tentative equivalent phrases or constructions in the target text. Similarly, machine translation produces many cases of wrong translations because of phraseology. The author pleads for more theoretical research taking into account the diversity of languages, and also for practical tools, of which an example is presented, the *IdiomSearch* tool, based on the automatic extraction of phraseology by means of a clustering algorithm.

One of the key issues of computational phraseology is to find which algorithms are best suited for the automatic or semi-automatic extraction of phrasemes, with possible differences according to the phraseme category. In chapter five, Alexander Wahl and Stefan Gries propose an algorithm for the automatic extraction of formulas, based on the progressive extension of bigrams according to the association strength. This methodology can also be used to help predict word sequences that young children will learn through language input. This shows that there is clearly a link between research on formulaic language and computational phraseology. While there is still room for improvement in the accuracy of the results obtained, this approach offers a comprehensive statistical discussion of the issues at stake in the extended bigram approach,

which is one of the promising avenues of research in the thorny issue of automatic extraction of phraseology.

Carlos Ramisch's contribution, in the next chapter, provides an overview of the complex interplay between phraseology and computational linguistics: for instance, natural language processing (NLP) uses existing phraseological resources, but on the other hand also contributes to the creation of new ones. As in the preceding two contributions, the focus of this chapter is on the extraction of phraseology, but from the point of view of the practical tools, as opposed to sometimes technical algorithms that are hard to reduplicate. The author describes the *mwetoolkit*, a combination of programs that may be combined with corpora in order to extract phrasemes by statistical scores. As the author points out, the very complete set of tools provided by the *mwetoolkit* might still be improved by means of a graphical interface and an implementation as a web application, but it already offers a concrete example of how the manipulation of comprehensive tools plays a crucial role in computational phraseology.

As pointed out by several contributions in this volume, huge linguistic corpora are necessary for gaining useful information in computational phraseology. In chapter seven, Peter Ďurčo therefore starts from a freely available collection of impressive size, the *Araneum corpora*. He goes on to discuss the respective advantages and drawbacks of comparable corpora, as opposed to monolingual and parallel corpora, for the analysis of phraseology. Crucially, the study shows that the recourse to corpora of unrelated texts is very useful for computational phraseology, provided that the corpora are compiled with the same methodology.

Along the phraseological cline, many weakly idiomatic combinations are traditionally regarded as *collocations*. In chapter eight, Marie-Claude L'Homme and Daphnée Azoulay shed some new light on one hitherto unexplored aspect of the behaviour of collocations: the difference between collocational patterns in general vs. specialised corpora. For the purpose of this study, they used collocates associated with 15 lexical items, extracted from a specialised corpus on the environment and a general corpus. Interestingly, some significant collocational behaviour can be found in both corpora, which may have consequences for future research and practical applications in terminology and lexicography.

The recourse to corpora in computational phraseology also poses the question of the choice between quantity and quality: is it preferable to use huge corpora of average quality, or to choose smaller ones with a higher degree of reliability? In chapter nine, Ruslan Mitkov and Shiva Taslimipoor are confronted with this problem in their search for translation equivalents of verb-noun collocations across comparable corpora (English and Spanish). This type of study is also of central importance to the research on machine translation, as there is also a debate between the *big data* approach and a more finely tuned selection of corpora. Their results actually show that both aspects should be taken into account.

The extraction of collocations from corpora has been investigated with several methodologies and statistical scores, but the question of the significance of the scores is a complex one, which is examined by Michael Oakes in the next chapter. By weighing the respective advantages and drawbacks of the currently used statistical scores, he underlines the relationship between the frequency, distribution and statistical scores. Giving a measure of the collocational strength is an additional issue, because the raw statistical results do not necessarily imply a gradation in the strength of the association.

In addition to the complex statistical framework and the variety of possible scores for the automatic extraction of collocations, a recurrent theme in computational phraseology is the possible relationship between syntactic structure and collocation extraction. In chapter eleven, Eric Wehrli, Violeta Seretan and Luka Nerima show that parsing is, on the one hand, beneficial to collocation extraction and that the latter can also be useful for improving parsing. In other words, the identification of collocations and syntactic parsing are claimed to be interrelated processes, which sheds more light on the interface between syntax and phraseology. Another original feature of this contribution is the inclusion of anaphora resolution in the extraction of collocations.

Phraseology as a whole is characterised by a high degree of frozenness, but there is also some kind of variation possible. The many systematic or contextual variants of phrasemes or multiword expressions have been thoroughly investigated in the literature, but the question remains what is the best description of those variations and to which level of description they are linked, such as grammar or statistical association. In chapter twelve, Luigi Squillante analyses the variation of multiword expressions in the case of Italian, and reaches the conclusion that the recourse to linguistic corpora and to grammatical principles offers a better methodology for describing this phenomenon.

In recent years, there have been many contacts between phraseology and another major theoretical approach to language in which idioms play an important role, namely construction grammar. As constructions are defined as (partly arbitrary) pairings of form and meaning, at various levels of abstraction, they also include phrasemes, but construction grammar offers fresh insights into the complex interplay between syntax and idiomaticity. *Collostructions* (a portmanteau word, from *collocations* and *constructions*) are particularly interesting at the crossroads of phraseology and construction grammar. In chapter thirteen, Anatol Stefanowitsch and Susanne Flach expand on collostructional analysis by starting from the patterns [too ADJ to V] and [ADJ enough to V].

Another way of looking at the interplay between constructions and phraseology, is to consider that what is at stake is a complex series of frozen lexical building blocks and of syntactic patterns. Using tools and corpora developed at the Institute for the German language in Mannheim, Kathrin Steyer shows in chapter fourteen that linguistic creativity in multiword expression is actually rooted in a number of syntactic patterns. She also demonstrates that a corpus-based or corpus-driven approach to computational phraseology, even though it relies on huge collections of linguistic data, must always be refined at the light of an appropriate selection of the relevant results.

Patrick Hanks addresses in chapter fifteen, the central issue of meaning and phraseology. If we claim that phraseology plays such an important role in language, there must be a way of connecting it to a semantic theory and, from a practical point of view, of explaining the meaning of words (partly) by phraseology. As lexical semantics turns out to be inconceivable without recourse to diverse contexts and preferred patterns, Hanks claims that the use of large electronic corpora will make it possible to map recurrent patterns of phraseology onto prototypical or stereotypical beliefs about meanings. In other words, phraseology may very well serve as a crucial link between semantics and language use.

In the last chapter, Shiva Taslimipoor, Gloria Corpas Pastor and Omid Rohanian present the results of a new methodology designed for establishing discriminative semantic differences.

This volume therefore concludes with a central issue in present-day and future work on computational phraseology, namely the complex links between phraseological associations and semantic ones. Indeed, Taslimipoor et al. reach signifiant results for semantic discrimination by having recourse to a number of techniques used in corpus linguistics (association scores, frequency on huge linguistic corpora) but also vector models and a knowledge-based ontology. Crucially, they show that, to some extent, phraseological association (as in the case of collocations) also contributes to the semantic network of words.

References

Colson, Jean-Pierre. 2003. Corpus Linguistics and Phraseological Statistics: a Few Hypotheses and Examples. In: *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zu Methodologie und Kulturspezifik der Phraseologie*, ed. by Harald Burger, Annelies Häcki Buhofer and Gertrud Gréciano, 47-59. Baltmannsweiler: Schneider Verlag.

Granger, Sylviane and Fanny Meunier (eds). 2008. *Phraseology. An interdisciplinary perspective*. Amsterdam / Philadelphia: John Benjamins.

Gries, Stefan Th. 2013. "50-something years of work on collocations. What is or should be next …" *International Journal of Corpus Linguistics* 18:137-165.

Heid, Ulrich. 2008. Computational phraseology: An overview. In: *Phraseology. An interdisciplinary perspective*, ed. by Sylviane Granger and Fanny Meunier, 337-360. Amsterdam / Philadelphia: John Benjamins.

Monti, Joanna; Seretan, Violeta; Corpas Pastor, Gloria and Ruslan Mitkov. 2018. Multiword units in machine translation and translation technology. *Multiword Units in Machine Translation and Translation Technology*, (Current Issues in Linguistic Theory, 341). Amsterdam and Philadelphia: John Benjamins. 1-37.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.